

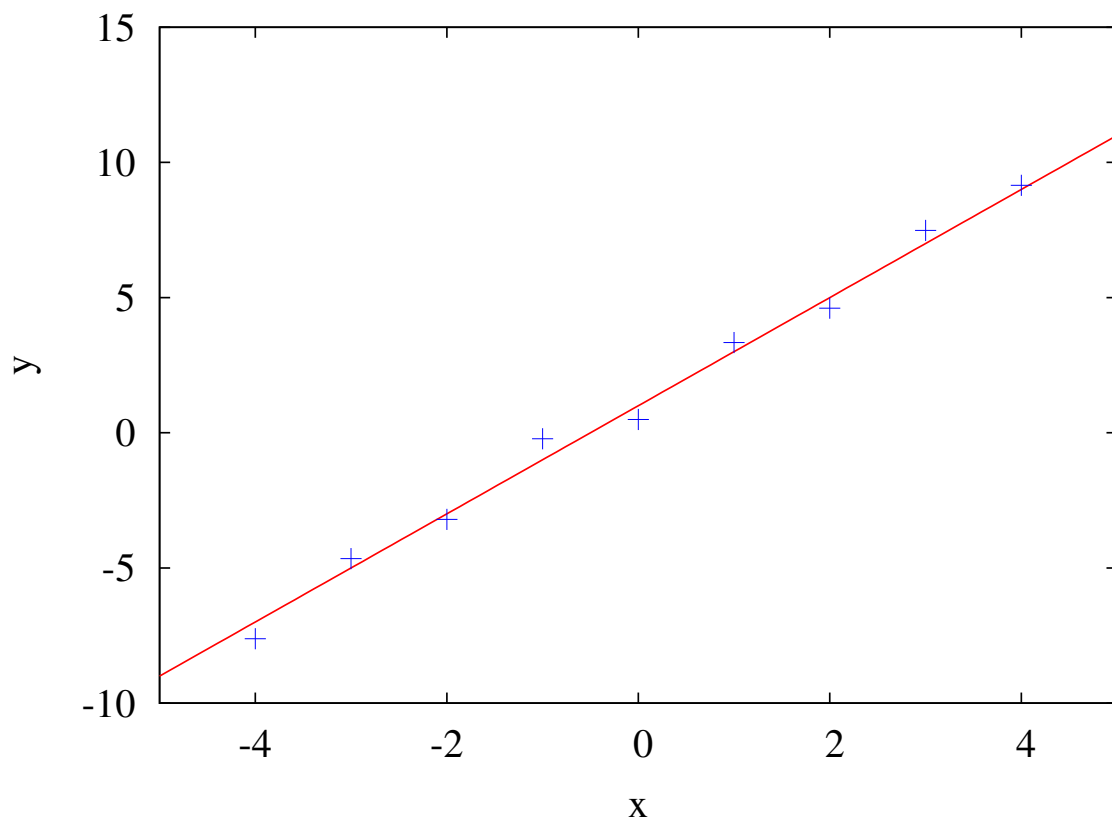
Nonlinear Approximation and Optimization

The problem with nonlinear optimization is local minima.

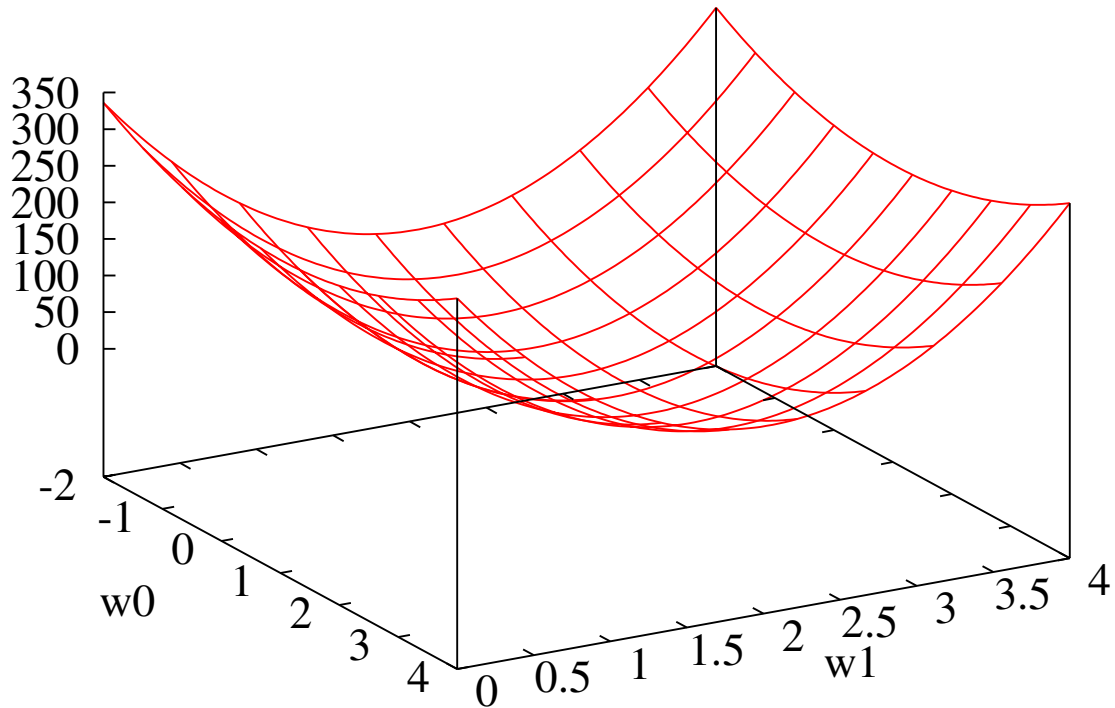
The following graphs illustrate approximating two functions

1. a noisy linear function $1 + 2x$ by $w_0 + w_1x$
2. a nonlinear function $2.5 \sin(1.5x)$ by $w_1 \sin(w_2 x)$.

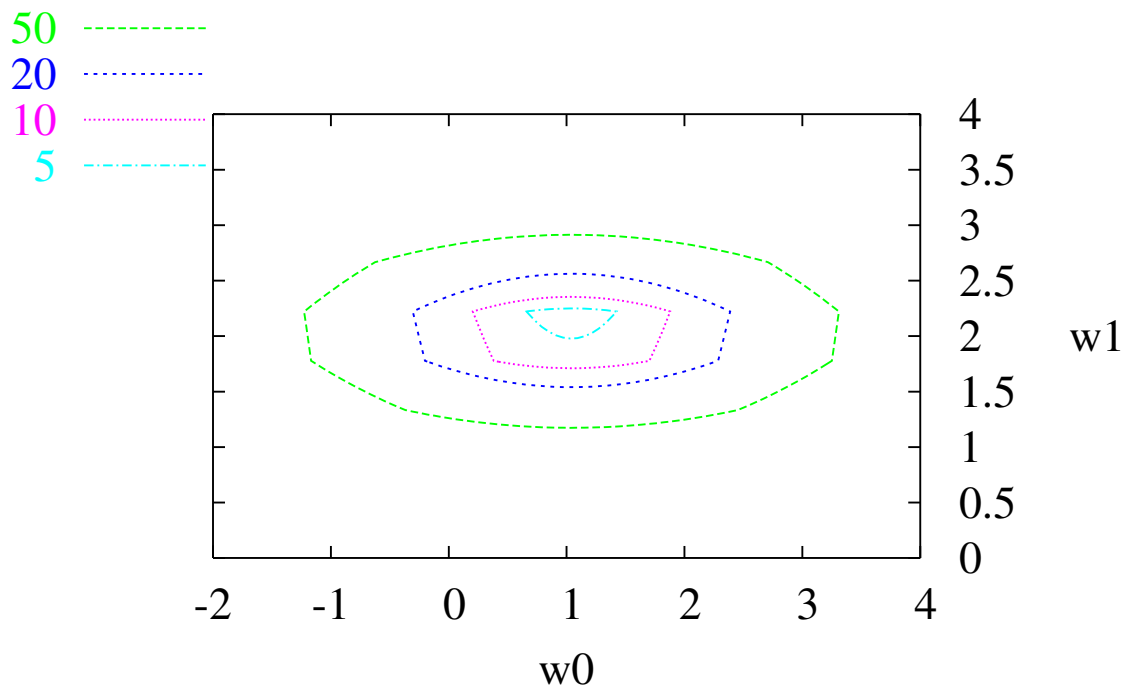
9 points from a noisy $1 + 2x$



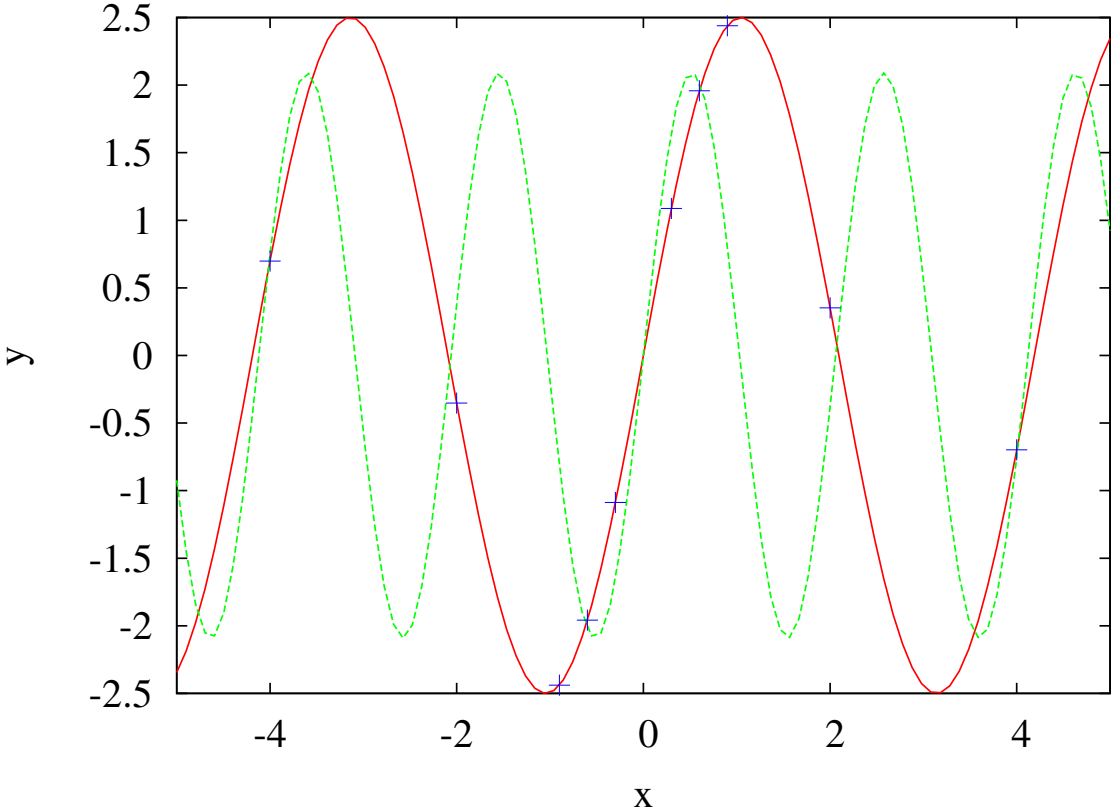
Error surface for weight values



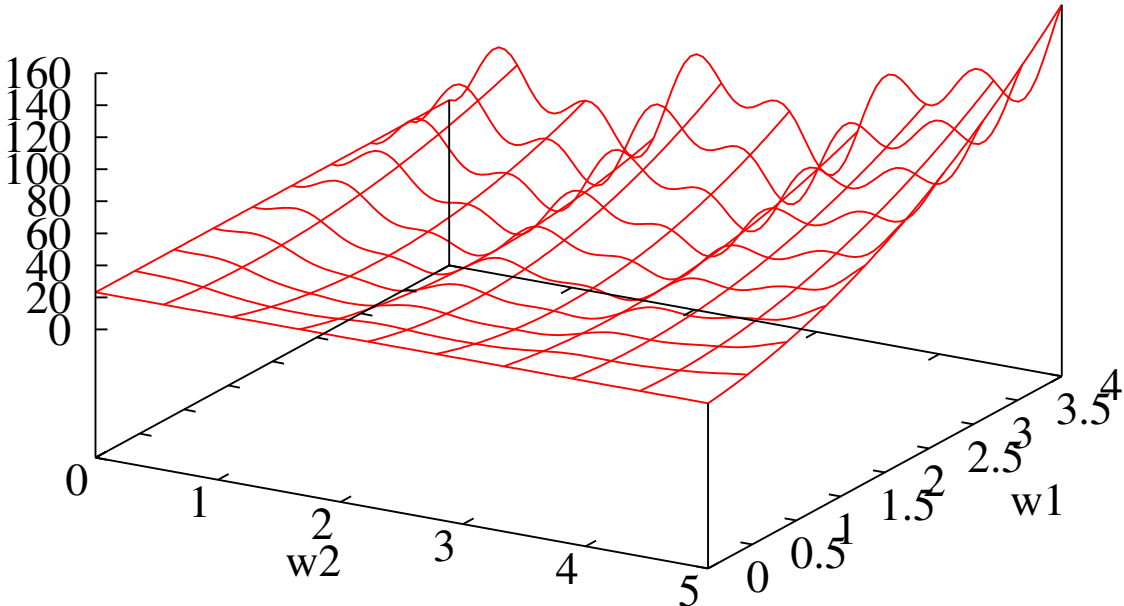
Error contours for weight values



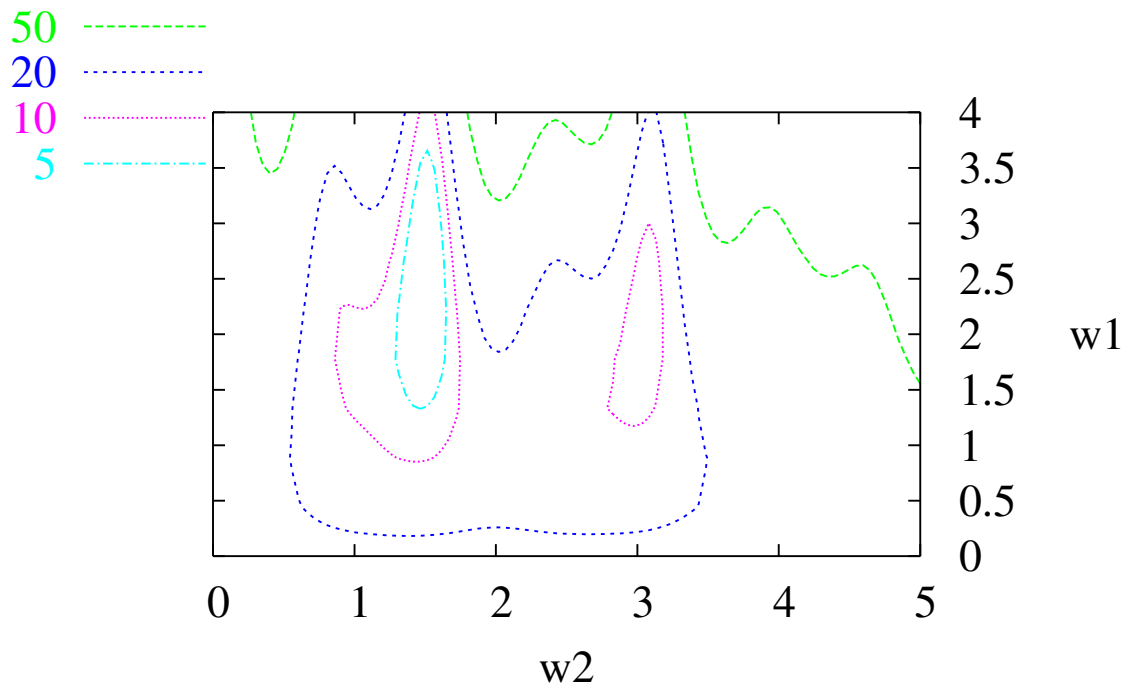
10 points from $2.5 \sin(1.5x)$



Error surface for weight values



Error contours for weight values



First Derivative Minimization

Let $\mathbf{g}(\mathbf{w})$ be the *gradient*, the partial derivatives of the error with respect to the weights.

$$\mathbf{g}(\mathbf{w}) = \nabla E(\mathbf{w}) = \frac{\partial E}{\partial \mathbf{w}} = \left[\frac{\partial E}{\partial w_1} \cdots \frac{\partial E}{\partial w_n} \right]$$

We can approximate the error near \mathbf{w} as:

$$\begin{aligned} E(\mathbf{w} + \mathbf{z}) &\approx E(\mathbf{w}) + \mathbf{z} \mathbf{g}(\mathbf{w})^T \\ &\approx E(\mathbf{w}) + \sum_{i=1}^n z_i \frac{\partial E}{\partial w_i} \end{aligned}$$

\mathbf{g} is the direction that increases error.
To decrease error, move in direction $-\mathbf{g}$.

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \mathbf{g}$$

η is the *learning rate*, which should be carefully chosen. If too high, error might increase.

The method is called *gradient descent*.

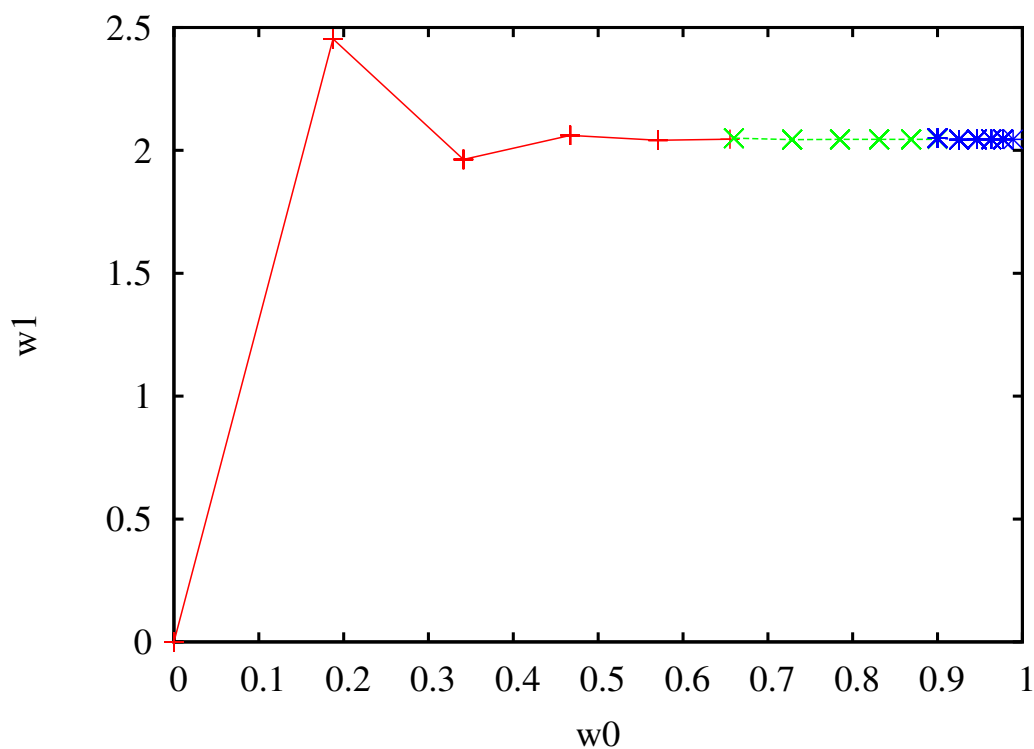
Disadvantages:

Global minimum might not be found.

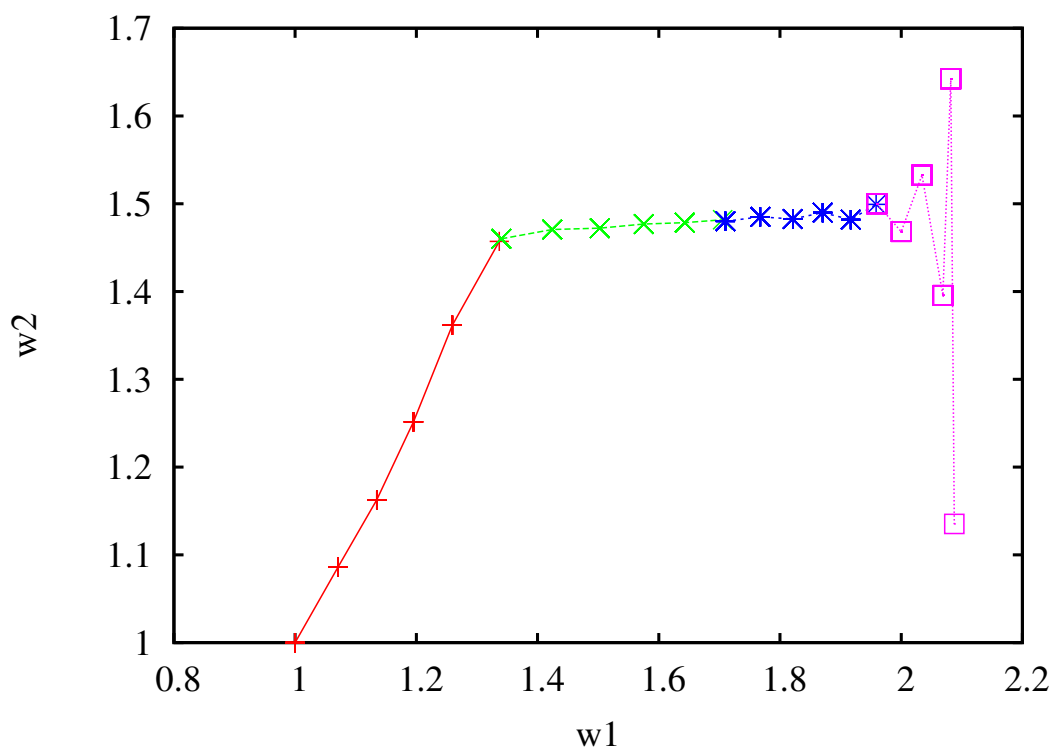
Inefficient when learning rate must be kept low.

Gradient descent on the noisy $1 + 2x$.

Squared error with $\eta = 0.01$.



Unstable gradient descent on $2.5 \sin(1.5x)$.
 Squared error with $\eta = 0.01$, starting at $(1, 1)$.



Deriving the Gradient

For an approximation $f_a(x, \mathbf{w})$, input-output pairs (x_i, y_i) , and squared error:

$$E(\mathbf{w}) = \sum_i (y_i - f_a(x_i, \mathbf{w}))^2$$

$$\frac{\partial E}{\partial \mathbf{w}} = \sum_i 2(y_i - f_a(x_i, \mathbf{w})) \frac{\partial f_a}{\partial \mathbf{w}}$$

$$\frac{\partial E}{\partial w_i} = \sum_i 2(y_i - f_a(x_i, \mathbf{w})) \frac{\partial f_a}{\partial w_i}$$

This method requires knowing the partial derivatives of the approximating function.

Second Derivative Minimization

The Hessian matrix is a matrix of second derivatives.

$$\mathbf{H}(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2 E}{\partial w_1^2} & \frac{\partial^2 E}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 E}{\partial w_1 \partial w_n} \\ \frac{\partial^2 E}{\partial w_2 \partial w_1} & \frac{\partial^2 E}{\partial w_2^2} & \cdots & \frac{\partial^2 E}{\partial w_2 \partial w_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 E}{\partial w_n \partial w_1} & \frac{\partial^2 E}{\partial w_n \partial w_2} & \cdots & \frac{\partial^2 E}{\partial w_n^2} \end{bmatrix}$$

We can approximate the error near \mathbf{w} as:

$$\begin{aligned} E(\mathbf{w} + \mathbf{z}) \\ \approx E(\mathbf{w}) + \mathbf{z} \mathbf{g}(\mathbf{w})^T + \frac{1}{2} \mathbf{z} \mathbf{H}(\mathbf{w}) \mathbf{z}^T \end{aligned}$$

If \mathbf{H} is constant and positive definite, the minimum is:

$$\mathbf{w} - \mathbf{H}^{-1} \mathbf{g}$$

See Chapter 8.1 for many methods based on using this approximation.